Final Synopsis

prior to the submission of the thesis

SOME CONTRIBUTIONS TO SURVIVAL ANALYSIS

to be submitted to

MANGALORE UNIVERSITY

for the award of the degree of

Doctor of Philosophy

in **STATISTICS**

by

SUMATHI K

under the supervision of

Prof. K. ARUNA RAO

Department of Post Graduate Studies and Research in Statistics

Mangalore University, Mangalagangotri 574199

Karnataka, India

August 2010

Introduction

Survival analysis is the branch of Statistics which deals with methods for analyzing the data regarding the time to occurrence of some events. The events could be the appearance of a disease or its recurrence, occurrence of death, failure of an electric bulb or breakdown of a machine. These events of interest are commonly known as lifetimes or failure times.

Although the need for the study of lifetime data frequently arises in health science, survival analysis does find its applications in various disciplines such as engineering and economics. There are several references on survival analysis which include Kalbfleisch and Prentice (1980), Miller (1981), Lawless (1982), Cox and Oakes (1984), Fleming and Harrington (1991), Andersen, Borgan, Gill and Keiding (1993), Klein and Moeschberger (1997) and Hougaard (2000).

Many times, during the study of lifetime data, the population may be heterogeneous and the investigator may also be interested in the factors associated with the event. In such situations, regression models can be used. For example, in a study on survival time for patients suffering from lung cancer, the relevant factors may be the age, the physical condition of the patients, the type of tumor and the time since diagnosis. These factors can be easily identified using regression models. The widely used regression model in survival analysis is the Cox (1972) proportional hazards model given by

$$h(t|X) = h_0(t)\exp(\beta' X) \qquad (1)$$

where T is a non-negative random variable denoting the lifetime of individuals, $h(t|X)$ is the hazard function of T given $X = (X_1, X_2, ..., X_p)$, a vector of unknown regressors associated with the individuals, $h_0(t)$ is an unknown baseline hazard common to all individuals and $\beta = (\beta_1, \beta_2, ..., \beta_p)$ is a vector of unknown regression coefficients.

The Cox (1972) proportional hazards model is based on the assumption that every individual in the population under study is susceptible to the adverse event under study and will eventually experience this event if the follow-up is sufficiently long. However there are situations when a proportion of the population under study do not experience this adverse event of interest. For example, the recent clinical trials indicate that a large group of patients are being cured of various chronic diseases such as cancer and epilepsy after a sufficient follow up. Thus a model for estimating the cured proportion was developed.

Boag (1949) developed a cure model to estimate the proportion of patients being cured among those who were receiving treatment for cancer of mouth and throat, cervix, uteri and breast. This model is known as the mixture cure model and is given by

$$S(t) = p + (1-p)S_u(t) \tag{2}$$

where p is the probability of a patient being cured on treatment, $S(t)$ is the survival function for the entire population of patients entering the clinical trial and $S_u(t)$ is the survival function for the uncured group of the population. This model was further modified by Berkson and Gage in 1952 and later studied extensively by several authors such as Kuk and Chen (1992), Laska and Meisner (1992), Maller and Zhou (1992, 1994, 1995, 1996) and Peng, Dear and Denham(1998).

Observing that the mixture model had some limitations when covariates were considered, Yakovlev et.al. (1993) developed an alternative to it and is known as the promotion time model or the bounded cumulative hazard (BCH) model given by

$$S(t) = \exp(-\theta F_u(t)) \tag{3}$$

The BCH model was developed by considering the patients suffering from cancer entering a clinical trial and was based on the assumption that after treatment a patient is left with N cancer cells capable of metastasizing. The process in which the cancer cells move to various parts of the human body, grow rapidly and replace the normal tissue is known as metastasis. The random variable N is assumed to follow a Poisson distribution with mean θ. If $t_k, k = 1,2,...N$, denotes the time for the k^{th} metastatic tumor cell to produce detectable cancer, then the time to relapse of cancer in the patient is given by $T = \min\{t_1, t_2, ..., t_N\}$. The variable T is observable. Conditional on N, it is assumed that t_k's are independent and identically distributed (i.i.d.) with cumulative distribution function $F_u(t)$, survival function $S_u(t)$ and probability density function $f_u(t)$, where the suffix u denotes the uncured group. The survival function $S(t)$ for the entire population of patients entering the clinical trial is given by

$S(t)$ = P[There is no detectable cancer by time t]

$$= P[N = 0] + P[t_1 > t, t_2 > t, ..., t_N > t \mid N \geq 1]$$

$$= \exp(-\theta) + \sum_{N=1}^{\infty} \left(S_u(t)\right)^N \frac{\exp(-\theta)\theta^N}{N!}$$

$$= \exp(-\theta F_u(t))$$

Here, the cured proportion p is given by $\lim_{t\to\infty} S(t) = \exp(-\theta)$.

The BCH model (3), also known as the promotion time cure model or the parametric cure model, can be written as a mixture model as follows.

$$S(t) = \exp(-\theta) + \exp(-\theta)(\exp(S_u(t)\theta) - 1)$$

$$= \exp(-\theta) + (1 - \exp(-\theta)) \frac{\exp(-\theta F_u(t)) - \exp(-\theta)}{(1 - \exp(-\theta))} \quad (4)$$

The survival function for the uncured population is given by

$$S_u(t) = P[T > t | N \geq 1]$$

$$= \frac{\exp(-\theta F_u(t)) - \exp(-\theta)}{(1 - \exp(-\theta))} \quad (5)$$

Thus every model defined by the BCH model (3) can be expressed as a mixture model with cured proportion $p = \exp(-\theta)$. In the presence of covariates X, the canonical link function $\theta = \exp(\beta' X)$ can be used. The literature available on BCH model is only in the Bayesian context. The details regarding this model can be found in Ibrahim, Chen and Sinha (2001).

Motivation for our study

The mixture cure model and the bounded cumulative hazard model are applicable only when the study variable is the time to occurrence of the adverse event. They do not consider the possibilities of the number of times the event recurs. In diseases like epilepsy or renal failures, the interest would be in determining the number of times the disease has recurred rather than the time intervals at which it recurred even after the patient is cured on treatment.

Clayton (1994) has discussed the analysis of recurrent event count data and the model proposed therein was based on the Poisson process. Although he brings out the connection between this model and the Cox proportional hazard model, Clayton does not address the issue of cure models for recurrent event count data. This motivated us to propose an inflated Poisson distribution (inflated at zero) as a cure model for recurrent event count data so as to enable the medical fraternity in estimating the cured fraction of the study population.

Overview of the thesis:

The thesis describes the newly proposed cure model for recurrent event count data and studies its inferential properties. It focuses on estimating the parameters in the model and testing for the cured proportion for recurrent events. An overview of the thesis is given below.

Chapter 1 gives a brief introduction to the thesis. The motivation for the study and chapter-wise summary of the entire thesis will also be discussed in brief.

Chapter 2 introduces the cure model and its applications in survival analysis. It discusses the two basic cure models viz., the mixture cure model and the BCH model and also a model based on Box-Cox (1964) transformation developed by Yin and Ibrahim (2005). An extension of a univariate to a bivariate set-up is possible through a function called copula. Therefore the chapter also provides a brief discussion on copulas and a short review on multivariate cure models. Finally, the new cure model for recurrent event count data is proposed. A brief description of the proposed model is given below.

Consider diseases like epilepsy, asthma and urinary tract infection, where the variable of interest is the number of times the disease recurs during the follow up period after the patient is said to be cured on treatment. In such cases, the factors associated with the response variable can be modeled using a discrete distribution. Motivated by the work of Clayton (1994), we observed that when the recurrent event data is modeled through Poisson distribution, the associated cure model is the inflated Poisson distribution. To simplify the situation, we assumed that the duration of follow up is the same for all the patients and there are no missing observations, that is, there are no patients who are lost to follow up. Let Y be the random variable denoting the number of times the event of interest occurs. Then the proposed new cure model for recurrent event count data based on a zero inflated Poisson (ZIP) model with probability mass function (p.m.f.) given by

$$P[Y=y]=\begin{cases} p+(1-p)\exp(-\lambda), \text{when } y=0 \\ (1-p)\exp(-\lambda)\dfrac{\lambda^{y}}{y!}, \text{when } y=1,2,... \end{cases} \tag{6}$$

Here p denotes the proportion of patients cured in the population. The covariates or the factors associated with the recurrent event can be linked to the mean parameter λ or to the inflate parameter p through the link functions

$$logit\ p_i = \log\left(\frac{p_i}{(1-p_i)}\right) = \alpha_0 + \alpha_1 z_{i1} + \ldots + \alpha_k z_{ik} \text{ and } log\ \lambda_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \quad (7)$$

where x_i and z_i are the covariate vectors and p_i and λ_i are parameters associated with the i^{th} individual.

The utility of the proposed model in real life situations is also discussed. **A paper based on a part of this chapter was presented at the Joint Statistical meeting and International Indian Statistical Association (IISA) Conference on Statistics, Probability and Related Areas organized by the Department of Statistics, Cochin University of Science and Technology (CUSAT), Cochin from January 2-5, 2007 and a paper based on the chapter has been published in the electronic indexed journal INTERSTAT in November 2008 issue.**

Chapter 3 provides a brief review of literature on zero inflated models (ZIM). This chapter has been included since the newly proposed cure model is based on a zero inflated distribution. The ZIM is defined as follows.

Consider a discrete random variable (r.v.) X taking at most a countable number of values $x = 0, 1, 2, \ldots$, with probability mass function (p.m.f) $P[X=x|\ \theta] = g(x, \theta)$ where the parameter $\theta \in \Omega \subset \mathbf{R}^s$. Define a new family of random variables having p.m.f $f(x, \theta, p)$ such that

$$f(x,\theta,p) = \begin{cases} p + (1-p)g(0,\theta), & x = 0 \\ (1-p)g(x,\theta), & x = 1,2,3,\ldots \end{cases} \quad (8)$$

$0 < p < 1,\ \theta \in \Omega$.

The r.v. X ~ ZIM (θ, g) or X ~ ZIM (θ) if there is no ambiguity regarding g. The p.m.f given by (8) is valid even when θ is vector valued. The p.m.f. given by (8) is a mixture distribution of two random variables, say X_0 degenerate at 0 and X_1 having p.m.f g in proportions p and (1-p) respectively. Some of the distributions that are used to obtain the inflated models are the Poisson, binomial, beta binomial, negative binomial, power series, generalized negative binomial and generalized Poisson. It may be noted that our new cure model (6) belongs to the family represented by (8). The origin of inflated distributions can be traced back to Neymann (1939) and Feller (1943). The work by Lambert (1992) on zero inflated Poisson (ZIP) regression in connection with the manufacturing defects popularized the use of regression models for the inflated count distribution.

With an introduction to ZIM and its applications, chapter 3 addresses the regularity conditions and the theoretical issues concerning the existence, uniqueness, consistency and asymptotic normality of maximum likelihood estimators and also the computational difficulties in solving the maximum likelihood equations of zero inflated distributions. A data set from dental epidemiological study has been analysed using ZIP, zero inflated negative binomial (ZINB) as well as zero inflated generalized Poisson (ZIGP) regression models.

A paper based on chapter 3 has been published in the electronic indexed journal INTERSTAT in August 2009 issue. Based on the same chapter, a paper discussing the difficulties in fitting a ZIP as well as a ZINB model to a data set has been presented at the International Conference on Statistics, Probability, Operations Research, Computer Science and Allied Areas organized by the Department of Statistics, Andhra University, Visakhapatnam from January 4-8, 2010. Based on the same data set, a comparison was made between a ZIP, ZINB and a ZIGP model and was presented as a paper at the East Asia Regional Biometric Conference & X Biennial Conference of International Biometric Society (Indian Region) held at Manipal University, Manipal on February 12th and 13th , 2010.

In **chapter 4**, the method of estimation of parameters in the proposed cure model (6) for recurrent event count data is described. The regularity conditions of Gan (1999) have been verified for our model. The mean square error (MSE) of the maximum likelihood estimator (MLE) of the cured proportion p as well as that of the mean parameter λ has been derived to the order $O\left(\frac{1}{n^2}\right)$. The covariance term is also derived. Fourteen tests are proposed for testing the cured proportion. Their asymptotic results have also been stated and proved. The covariates are not considered. The first three tests are the likelihood ratio test, the Wald test and the score test. The remaining tests are the locally most powerful tests and the perturbed versions of the Wald and the score tests. Simulations carried out to check the normal approximation of the test statistics are presented.

The simulation configurations are as follows:

The level of significance $\alpha = 0.05$

Sample size n = 20, 40, 80, 100, 200, 400

p = 0.1, 0.3, 0.5, 0.7, 0.9

λ= 1, 3, 5, 7, 9

The number of simulations = 10000

The comparisons of the tests are made based on their type I error rates and powers.

A paper based on this chapter was presented at the VI International Symposium on Optimization and Statistics (ISOS- 2008) organized by the Department of Statistics and Operations Research, Aligarh Muslim University, Aligarh from December 29-31, 2008 and has been published in the special issue of the proceedings of ISOS-2008 of their journal "Optimisation and Statistics", after being refereed.

Chapter 5 is an extension of the work discussed in chapter 4. In chapter 4, the work was based on data when there was no censoring. But in failure time models, it is very common that patients are lost to follow up and thus we encounter censored data, especially in long term follow up studies. Clayton (1994) suggests that when there is censoring, the censored observations can be treated as missing observations and the procedure adopted to handle the missing observations can be applied to the case of censored situations. **But we suggest an alternative procedure when there is censoring. Our procedure is based on the assumption that information is available on the censoring times as well as the number of recurrences of the disease till the time of censoring**. **Under this assumption, for the n individuals on follow up, the number of recurrences follow a zero inflated Poisson process with parameters (p_i, $\lambda \tilde{t}_i$), i = 1, 2, …, n, where $\tilde{T}_i$ is the random variable denoting the censoring time.** In chapter 5, we develop conditional tests for cured proportion when censoring times are known. The asymptotic normality of the conditional MLEs are proved. The chapter discusses thirteen tests proposed for testing the cured proportion, which are extensions of the tests proposed in chapter 4. The performance of the tests are examined using simulations. 20% censoring has been considered. The maximum follow up period was taken as T = 60 months. We considered the censoring time $\tilde{T}_i$ as a uniform random variable. The remaining simulation configurations are the same as those considered in chapter 4 of the thesis. The comparisons of the tests are again made based on their type I error rates and powers.

A paper based on this chapter was presented at the International Conference on Statistics, Probability, Operations Research, Computer Science and Allied Areas (IISA

& ISPS) organized by the Department of Statistics, Andhra University, Visakhapatnam from January 4-8, 2010.

Chapter 6 is an extension of the study discussed in chapter 4. In chapters 4 and 5, the tests were proposed for cured proportion when the data were uncensored and censored respectively, but without considering the covariates. However, in long term follow studies, it is a practice to record all the factors relevant to the event of interest. Therefore in chapter 6, the study carried out in chapter 4 has been extended to the case where covariates are involved. We considered one covariate only. The MLEs and the results based on their asymptotic normality for the parameter p and the regression coefficients are established. Thirteen tests for cured proportion are proposed. The covariates are related to the mean parameter λ through the log link function. Discrete as well as continuous covariates are considered. The simulation configurations remain the same as that of chapter 4. The comparisons of the tests are once again based on their type I error rates and powers.

Chapter 7 summarises the work of the entire thesis and gives directions for future research.

List of Publications :

1. Sumathi, K. and Rao, A.K (2010). Tests for Cured Proportion for Recurrent Event Data. *Optimisation and Statistics.*

2. Sumathi K and Rao A.K.(2009). On Estimation and Tests For Zero Inflated Regression Models. *INTERSTAT.*

3. Sumathi, K. and Rao, A.K (2008). Cure rate models – A Partial Review with an application to recurrent event or count data. *INTERSTAT.*

Conference Presentations:

1. Rao, K.A., and Sumathi K. (2010). Cure Models for Recurrent event data. *Presented at the East Asia Regional Biometric Conference & X Biennial Conference of International Biometric Society (Indian Region) held at Manipal University, Manipal on February 12th and 13th , 2010.*

2. Sumathi. K and Rao, K.A. (2010). The Zero-inflated Generalized Poisson Model and the Decayed, Missing and Filled Teeth (DMFT) index in dental epidemiology. *Presented at the East Asia Regional Biometric Conference & X Biennial Conference of International Biometric Society (Indian Region) held at Manipal University, Manipal on February 12th and 13th , 2010.*

3. Rao, K.A., and Sumathi K. (2010). Tests for cured proportions of recurrent event data under random censoring. *Presented at the International Conference on Statistics, Probability, Operations Research, Computer Science and Allied Areas (IISA & ISPS) organized by Department of Statistics, Andhra University, Visakhapatnam from January 4-8, 2010.*

4. Sumathi K. and Rao, K.A. (2010). On epidemiological models using zero inflated distributions: A case study of Cross-sectional data from a dental set up. *Presented at the International Conference on Statistics, Probability, Operations Research, Computer Science and Allied Areas (IISA & ISPS) organized by Department of Statistics, Andhra University, Visakhapatnam from January 4-8, 2010.*

5. Sumathi K and Rao, K.A.(2008). On Tests for cured proportion for recurrent event data. *Presented at VI International Symposium on Optimization and Statistics (ISOS- 2008) organized by Department of Statistics and Operations Research, Aligarh Muslim University, Aligarh December 29-31, 2008*

6. Sumathi. K and Rao, K.A. (2007). Cure Rate Models: A Review. *Presented at Joint Statistical meeting and International Indian Statistical Association (IISA) Conference on Statistics, Probability and Related Areas Organized by Department of Statistics, Cochin University of Science and Technology* (*CUSAT), Cochin, from January 2-5, 2007.*

References:

1. Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer-Verlag.

2. Berkson, J. and Gage, R. P. (1952). Survival curves for cancer patients following treatment. *Journal of the American Statistical Association* **47**, 501-515.

3. Boag,J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B* **11,** 15-44.

4. Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26**, 211-252.

5. Clayton,D.G.(1994). Some Approaches to the analysis of recurrent event data. *Statistical methods in medical research* **3**,244-262

6. Cox,D.R and Oakes, D. (1984). *Analysis of Survival Data.* UK: Chapman and Hall.

7. Cox,D.R (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

8. Feller, W. (1943). On a general class of contagious distributions. *Annals of Mathematical Statistics,* **16,** 319-329.

9. Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

10.Gan, N. (1999). *General zero inflated models and their applications*. Unpublished Ph.D. thesis, North Carolina State University, North Carolina.

11. Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer-Verlag.

12. Ibrahim, J.G., Chen, M.H. and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag.

13. Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

14. Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis.* New York: Springer-Verlag.

15. Kuk,A.Y.C, and Chen,C.H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79**, 531-541.

16. Lambert, D. (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics* **34**, 1-13.

17. Laska,E.M., and Meisner,M.J. (1992). Nonparametric estimation and testing in a cure rate model. *Biometrics* **48**, 1223-1234.

18. Lawless, J. F. (1982). Statistical Models and Methods for Lifetime Data. New York: Wiley

19. Maller, R. A. and Zhou, S. (1996) . *Survival Analysis with Long-Term Survivors.* New York: Wiley.

20. Maller, R. A. and Zhou, S. (1995). Testing for the presence of immune or cured individuals in censored survival data .*Biometrics* **51,** 1197-1205.

21. Maller ,R .A and Zhou,S. (1994). Testing for sufficient follow-up and outliers in survival data . *Journal of the American Statistical Association* **89,** 1499-1506.

22. Maller ,R .A and Zhou,S. (1992). Estimating the proportion of Immunes in a censored sample. *Biometrika* **79,** 731-739.

23. Miller, R.G.(1981). *Survival Analysis* , Chichester, U.K : Wiley .

24. Neymann, J. (1939). On a class of contagious distributions applicable in entomology and bacteriology, *Biometrika*, **30**, 11-15.

25. Peng,Y., Dear,K.B.G., and Denham,J.W. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine* **17**, 813-830.

26. Yakovlev A.Y., Asselain, B., Bardou, V.J., Fourquet, A., Hoang, T., Rochefediere, A., and Tsodikov, A.D (1993) . A Simple Stochastic Model of Tumor Recurrence and Its Applications to Data on pre-menopausal Breast Cancer. In *Biometrie et Analyse de Dormees Spatio – Temporelles* **12** (Eds. B. Asselain, M. Boniface, C. Duby, C. Lopez, J.P.Masson, and J.Tranchefort). Société Francaise de Biométrie, ENSA Renned, France, pp. 66-82.

27. Yin, G and Ibrahim , J.G. (2005). Cure rate models : a unified approach .*The Canadian Journal of statistics* **33**, 559-570

****

.